Graph Evolution via Social Diffusion Processes

Dijun Luo **Chris Ding** Heng Huang
University of Texas at Arlington

The University of Texas
ARLINGTON™

Outline

- Introduction
- Motivation
- Social Diffusion Processes
- Applications
- Experimental Results
- Conclusions

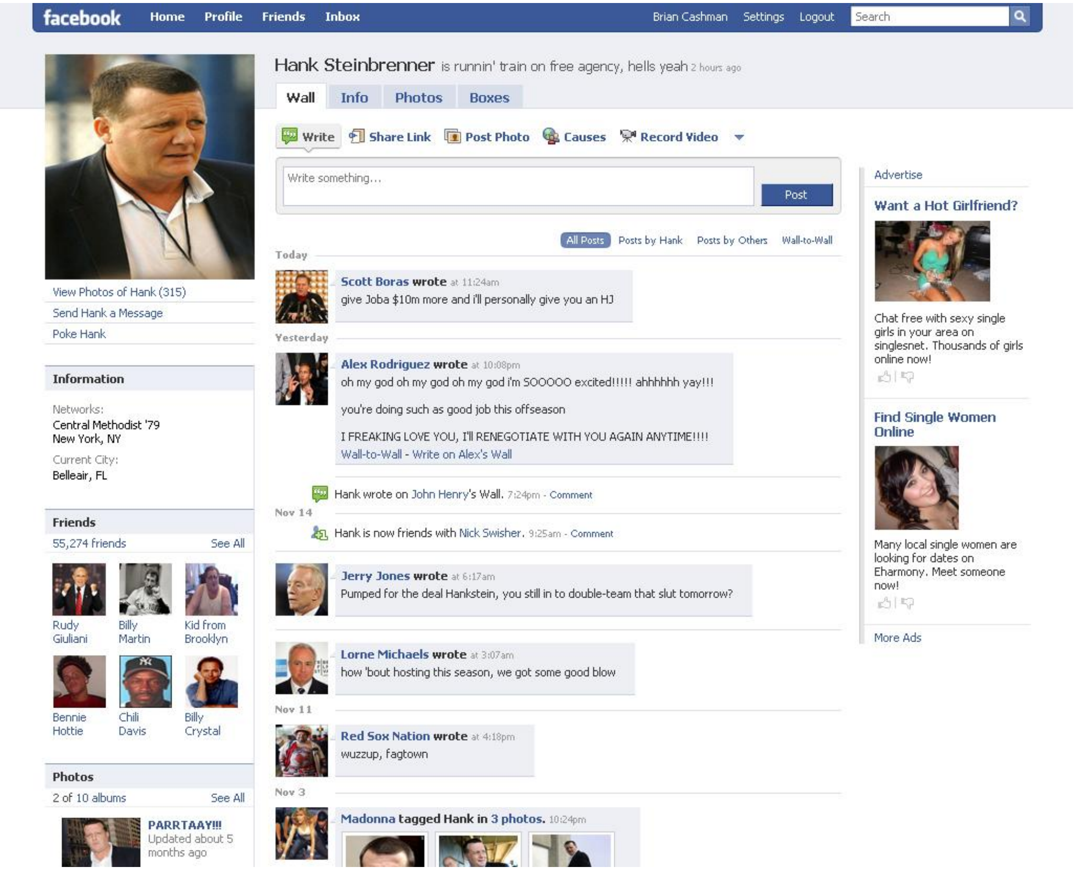
Introduction

- Graph-based clustering approaches are widely employed
  - Simple, easily to understand, good results [Shi-Malik1997, Ng et al 2001, Chan et al 1993]
  - Graph data are widely available
- Most of previous research focus on **static analysis** of graph
  - *Graph partition* seeks grouping using static optimization, cut edges between clusters
  - *Stochastic modeling* maximize the likelihood of a generative model on the graph.
- Our work present a novel **dynamic analysis** of graph data
  - Inspired by *Matthew effect*, a general phenomenon in nature and societies
  - Stronger connections become stronger
  - Expand and smooth social circles

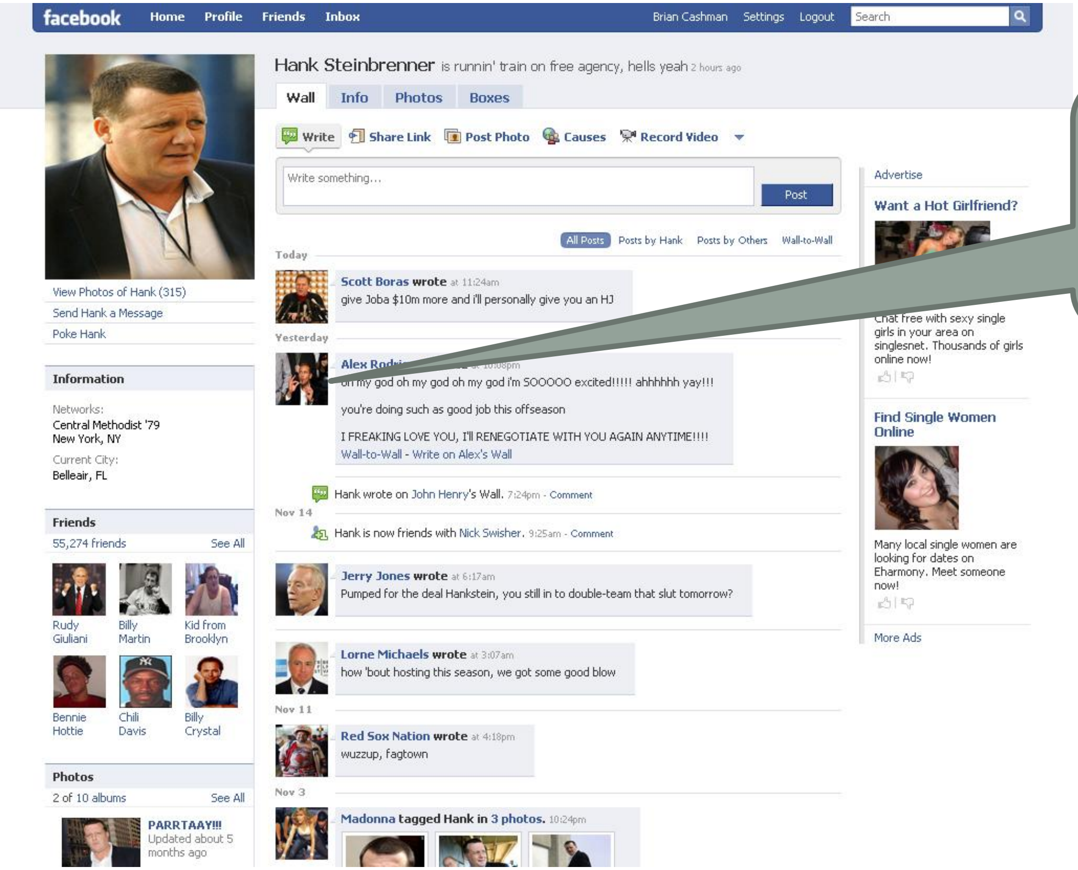
Motivation

- The relationship among people in a society changes in time
  - People are typically involved in many social events
  - E.g. meeting new friends, attending conferences like ECML here
    - The more we meet with each other in a conference, the more familiar we are
  - People will connect with each other using the connection, like meeting friends' friends
- Several observations
  - Two people with many common friends have a lot of chance to know each other
  - Two good friends have good chances to meet in the same social events, hence they know each more
- Social Diffusion Process
  - An analogue of the social relationship evolution

Motivation case study: Facebook

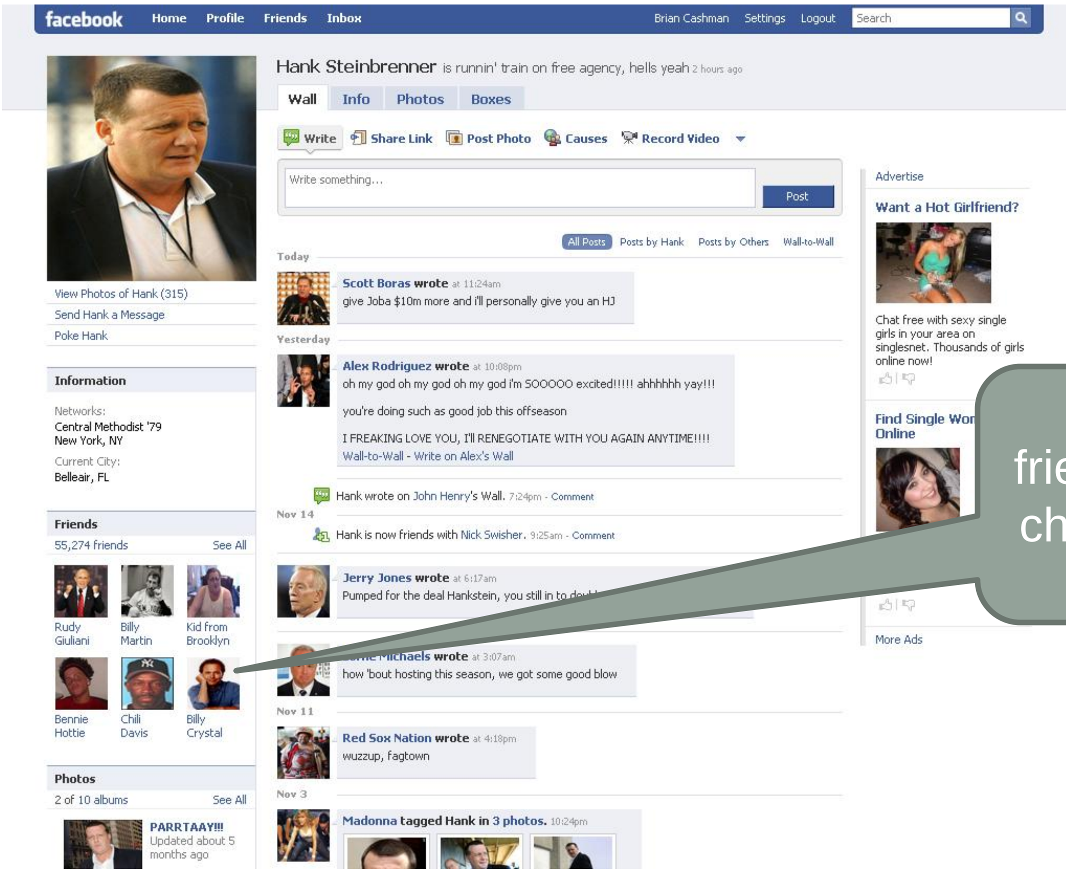


Motivation case study: Facebook



We will see the events of our friend's friends

Motivation case study: Facebook



More common friends means more chance to know the event

Social Diffusion Process

- Two friends set up a date. They meet.
- Two friends set up a date. One brings along a friend. The three of them meet.
- Two friends set up a date. Both friends bring along a friend each. The four of them meet.

There exist more processes. But these are the most fundamental processes. We consider them only in this work.

Social Diffusion Process

- Two friends set up a date. They meet.

- Two friends (A,B) set up a date. One (B) brings along a friend (C). The three of them meet.
  - A meets C

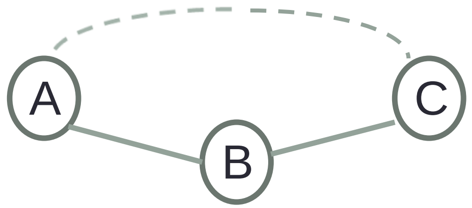
- Two friends (A,B) set up a date. Both friends bring along a friend [A brings C. B brings D]. The four of them meet.
  - A meets D
  - B meets C;
  - Most importantly, C meets D

- Diffusion: two person meet due to their friends' initiative

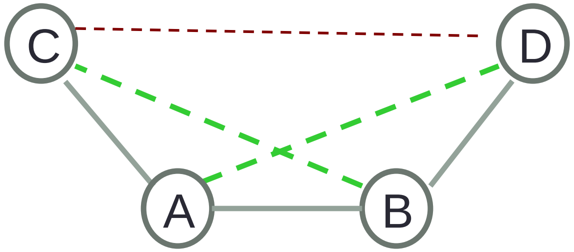
Social Diffusion Process

- Two friends set up a date. They meet.

- Two friends (A,B) set up a date. One (B) brings along a friend (C). The three of them meet.
  - A meets C (two person meet due to a common friend)



- Two friends (A,B) set up a date. Both friends bring along a friend [A brings C. B brings D]. The four of them meet.
  - A meets D (two person meet due to a common friend)
  - B meets C (two person meet due to a common friend)
  - Most importantly, C meets D (two person meet due to a friend's friend)



Social Diffusion Process

- Three social events
  - Date(v_i, v_j) : social players v_i and v_j initial a dating
  - Bring(v_i, v_k): social play v_i bring v_k when dating with some other player v_j
  - Meet(v_i, v_j): : social players v_i and v_j meet in a social event
- Rules

Two friends setup date. They meet

$$\text{Rule 1:} \quad \mathbf{Date}(v_i, v_j) \Rightarrow \mathbf{Meet}(v_i, v_j)$$

Two friends setup date. One brings along a friend. They meet.

$$\text{Rule 2:} \quad \left.\begin{array}{l} \mathbf{Date}(v_i, v_j) \\ \mathbf{Bring}(v_i, v_k) \end{array}\right\} \Rightarrow \mathbf{Meet}(v_j, v_k)$$

Two friends setup date. Both bring along a friend. They meet.

$$\text{Rule 3:} \quad \left.\begin{array}{l} \mathbf{Date}(v_i, v_j) \\ \mathbf{Bring}(v_i, v_k) \\ \mathbf{Bring}(v_j, v_l) \end{array}\right\} \Rightarrow \mathbf{Meet}(v_k, v_l)$$

Social Diffusion Process

- Assume we want to date with some one on the wedding of Royal wedding for William and Kate, who are we going to date?
  - We will bring important friends
- Observations
  - We will choose different level of friends to attend a different events
  - The **bring-friend** action should have a threshold

Social Diffusion Process

- Social Diffusion Process is a process as follows
  - (1) Choose a threshold $t \sim U(0, \mu)$
  - (2) Date(v_i, v_j) happens if $w_{ij} > t$
  - (3) For any k, l
    - Bring(v_i, v_k) and Bring(v_j, v_l) happen with probability

$$p(i,k,t) = \begin{cases} \frac{1}{|\mathcal{N}_{i,t}|} & \text{if } \; v_k \in \mathcal{N}_{i,t} \\ 0 & \text{otherwise} \end{cases}$$

$$p(j,l,t) = \begin{cases} \frac{1}{|\mathcal{N}_{j,t}|} & \text{if } \; v_k \in \mathcal{N}_{j,t} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{N}_{i,t} = \{q : W_{iq} \geq t\}, \mathcal{N}_{j,t} = \{q : W_{jq} \geq t\}$$

$$\text{if } \mathbf{Meet}(v_p, v_q),\ W_{pq} \leftarrow W_{pq} + \alpha\mu$$

Social Diffusion Process

- Social Diffusion Process is a process as follows
  - (1) Choose a threshold $t \sim U(0, \mu)$ ← Uniform distribution
  - (2) Date(v_i, v_j) happens if $w_{ij} > t$
  - (3) For any k, l
    - Bring(v_i, v_k) and Bring(v_j, v_l) happen with probability

$$p(i,k,t) = \begin{cases} \frac{1}{|\mathcal{N}_{i,t}|} & \text{if } \ v_k \in \mathcal{N}_{i,t} \\ 0 & \text{otherwise} \end{cases}$$

$$p(j,l,t) = \begin{cases} \frac{1}{|\mathcal{N}_{j,t}|} & \text{if } \ v_k \in \mathcal{N}_{j,t} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{N}_{i,t} = \{q : W_{iq} \geq t\}, \mathcal{N}_{j,t} = \{q : W_{jq} \geq t\}$$

$$\text{if } \mathbf{Meet}(v_p, v_q),\ W_{pq} \leftarrow W_{pq} + \alpha\mu$$

$$\mu = \max_{ij} W_{ij}$$

Diffusion constant (α)
Set to 1 in algorithm

Social Diffusion Process Model

Define thresholded graph adjacency matrix as

$$(A^t)_{ij} = \begin{cases} 1 & \text{if } W_{ij} \geq t \\ 0 & \text{otherwise} \end{cases}$$

Proportional constant
Set to 1 in algorithm

Case (1). **Date**(v_i, v_j). In this case the probability that they meet is

$$P(\textbf{Meet}(v_i, v_j)) = \delta (A^t)_{ij}.$$

Case (2). **Date**(v_i, v_k) and **Bring**(v_k, v_j). By definition $|\mathcal{N}_{k,t}| = \sum_j A^t_{jk} = d^t_k$, where d^t_k is the degree k in A^t. In this case,

$$\textit{random walk probability}: P_{k \to i} = \frac{A^t_{ki}}{d_k}$$

$$\begin{aligned} & P(\textbf{Meet}(v_i, v_j)) \\ & = \sum_k P(\textbf{Meet}(v_i, v_j) | \textbf{Date}(v_i, v_k), \textbf{Bring}(v_k, v_j)) \\ & = \sum_k \delta (A^t)_{ik} \frac{A^t_{jk}}{d_k} = \delta (A^t D^{-1} A^t)_{ij}, \end{aligned}$$

Social Diffusion Process

$$random\ walk\ \ probability: P_{k\to i} = \frac{A_{ki}^t}{d_k}$$

Case(3). **Date**(v_k, v_l), **Bring**(v_k, v_i), and **Bring**(v_l, v_j). Similar with case (2), we have

$$P(\mathbf{Meet}(v_i, v_j)) = \sum_{kl} \delta(A^t)_{kl} \frac{A_{ik}^t}{d_k} \frac{A_{jl}^t}{d_l}$$
$$= \delta(A^t D^{-1} A^t D^{-1} A^t)_{ij}.$$

By summing up the three cases, we have

$$P(\mathbf{Meet}(v_i, v_j))$$
$$= \delta A_{ij}^t + \delta(A^t D^{-1} A^t)_{ij} + \delta(A^t D^{-1} A^t D^{-1} A^t)_{ij}$$

Diffusion constant
Set to 1 in algorithm

$$\mathbb{E}(\Delta W_{ij})$$
$$= \alpha\mu\delta\left(A_{ij}^t + (A^t D^{-1} A^t)_{ij} + (A^t D^{-1} A^t D^{-1} A^t)_{ij}\right)$$
$$\triangleq \alpha\mu\delta M_{ij}^t.$$

$$\mu = \max_{ij} W_{ij}$$

Social Diffusion Process Algorithm

Algorithm 1 $\tilde{W}$ = **GraphEvolution**(W)

Input: Graph W

Output: Graph $\tilde{W}$

$\mu = \max_{ij} W_{ij}, \tilde{W} = \mathbf{0}$

for $i = 1 : T$ ← The only model parameter

$t = i\mu/T$

Calculate M^t using Eq. (5)

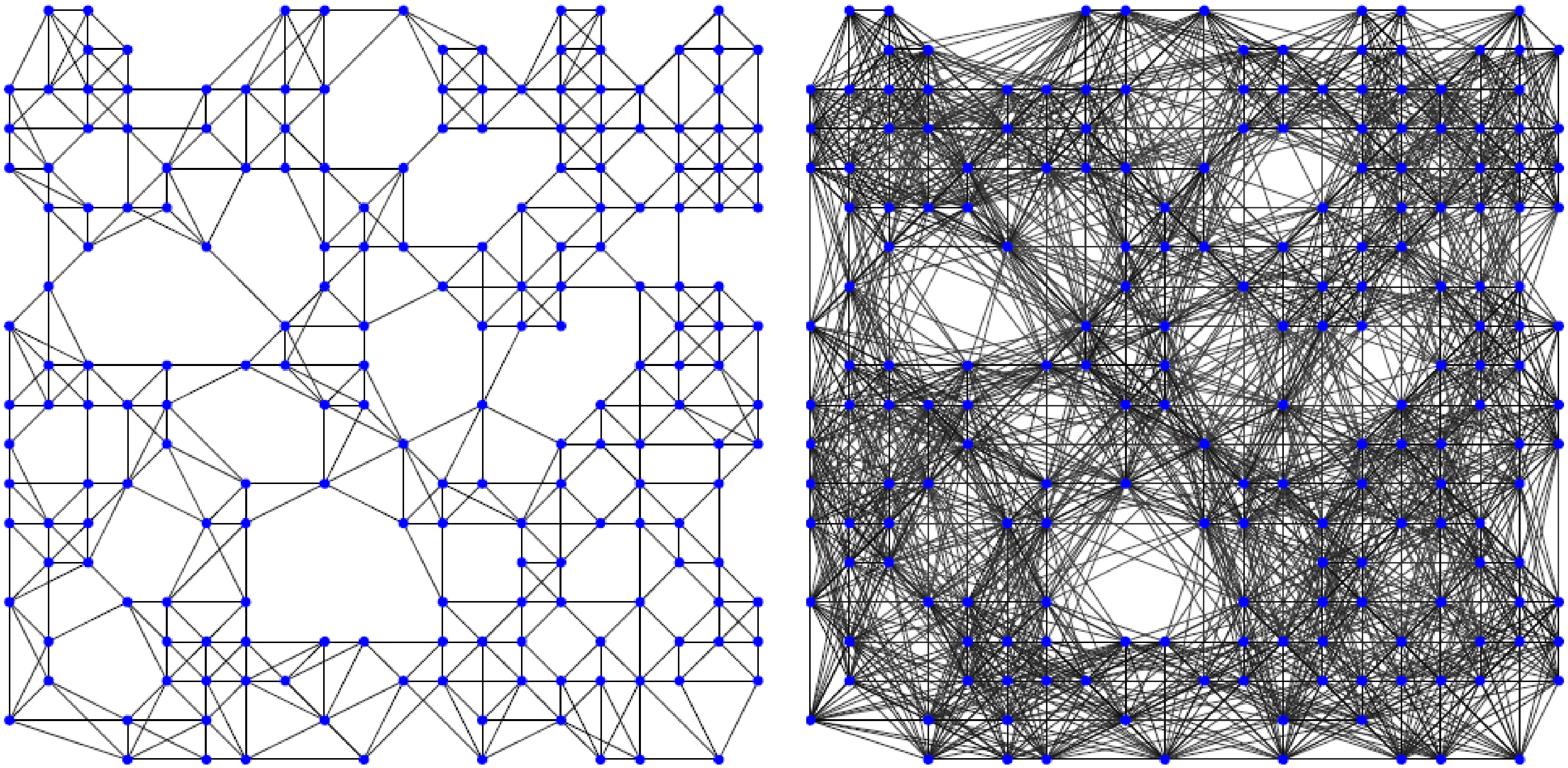
Normalize $M^t : M^t_{ij} \leftarrow M^t_{ij} / \sum_{i'j'} M^t_{i'j'}$

$\tilde{W} \leftarrow \tilde{W} + M^t$

end for

Output: $\tilde{W}$

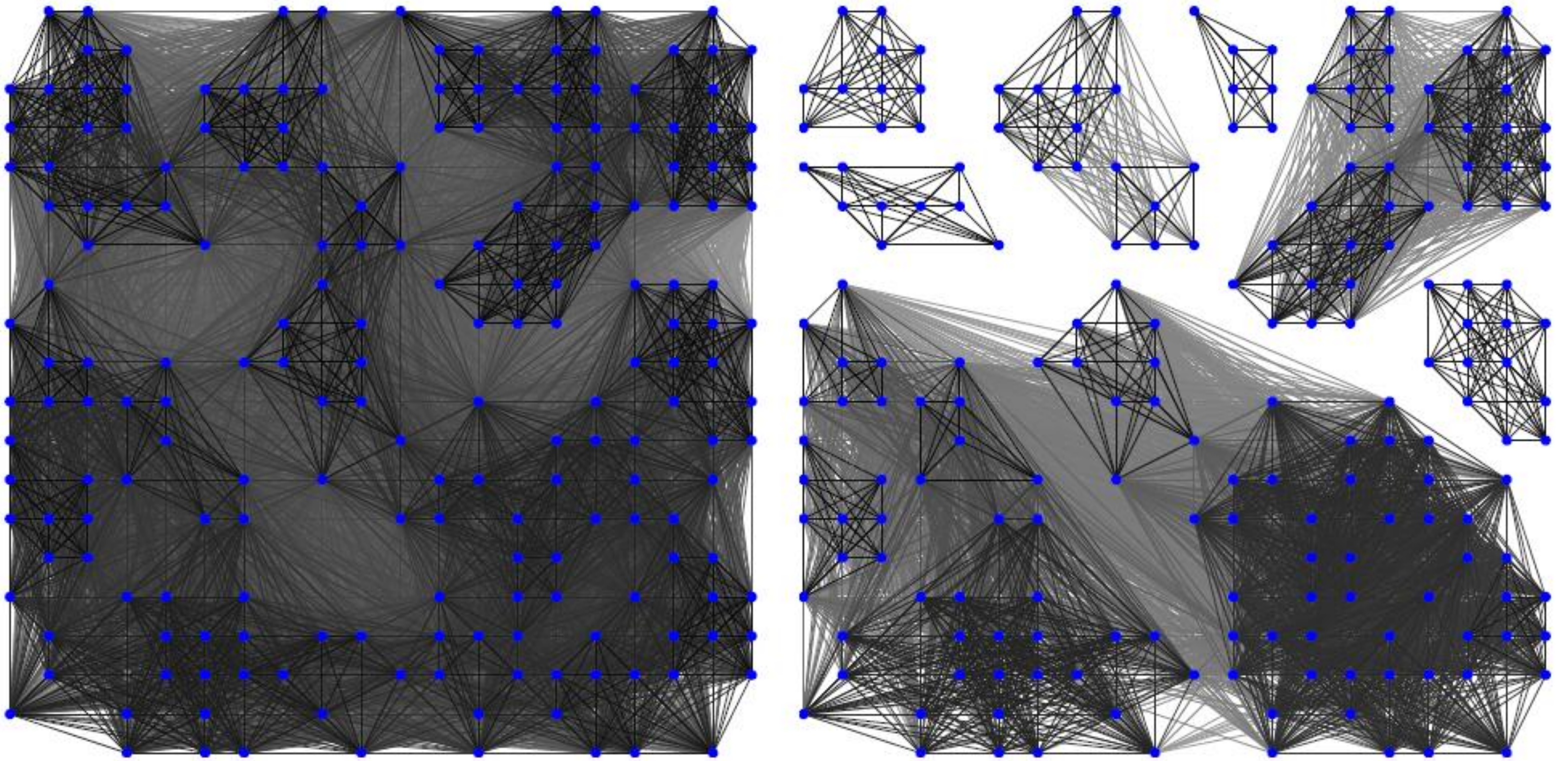
Social Diffusion Process: a simple case



(a) Initialization (b) 1st iteration

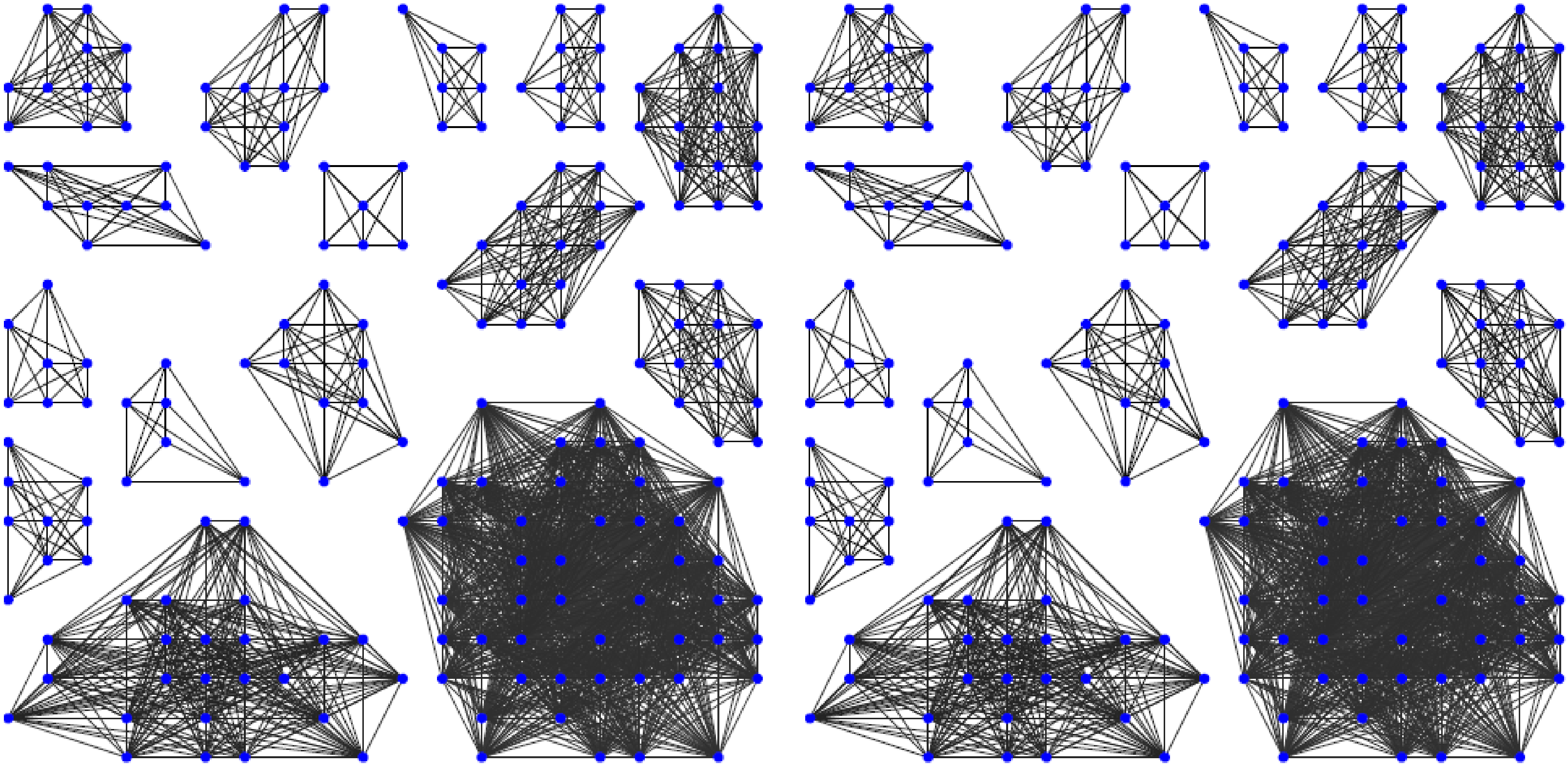
$$W \leftarrow \mathbf{GraphEvolution}(W)$$

Social Diffusion Process: a simple case



(c) 3rd iteration (d) 10th iteration

Social Diffusion Process: a simple case



(e) 15th iteration (f) 20th iteration

Applications

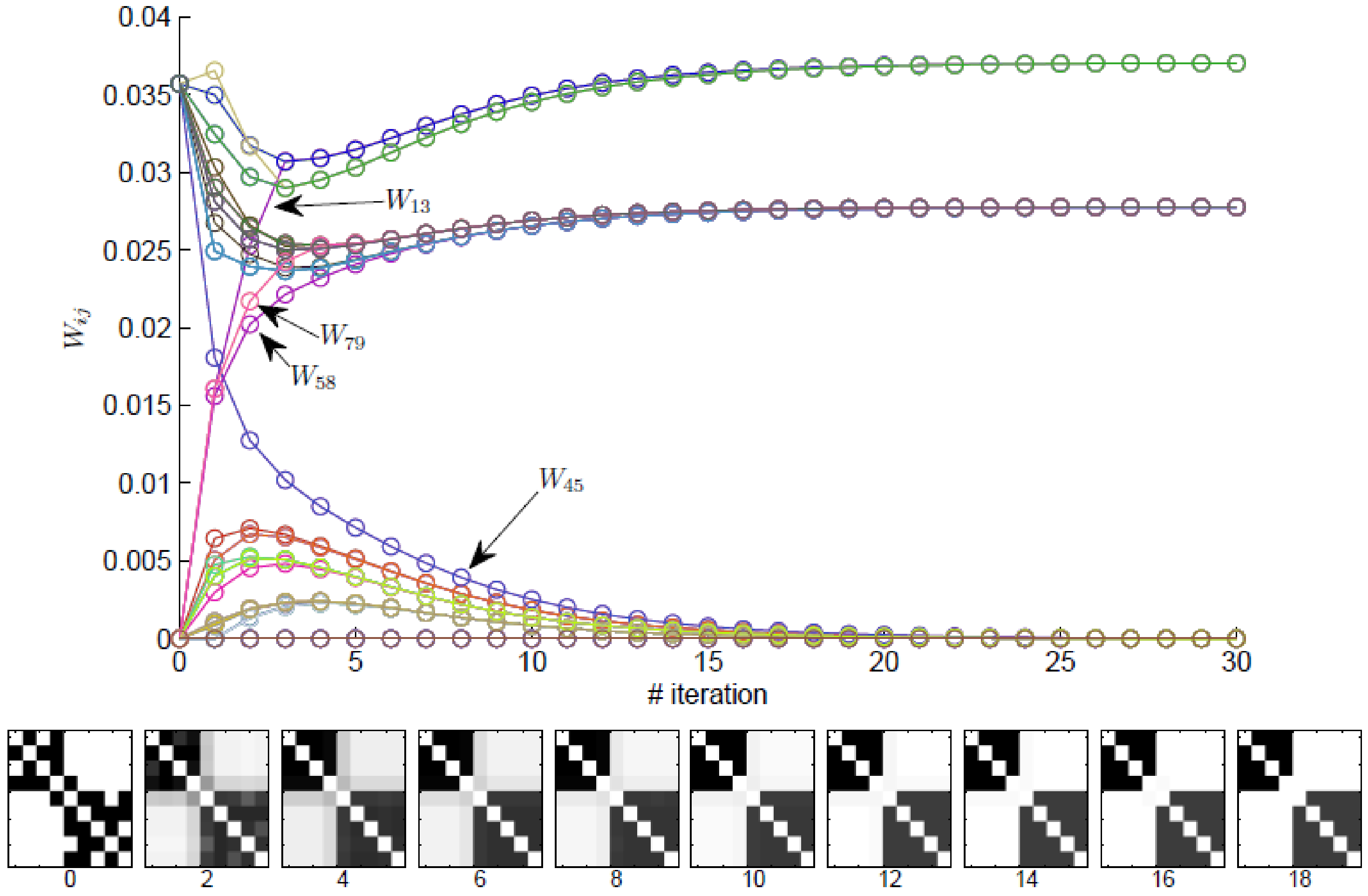
- Clustering
  - Grouping results can be derived when disconnected components are observed
- Preprocessing for other machine learning tasks
  - Our algorithm take a graph as input and a better graph as output
  - Can be used as preprocessing
  - Clustering, semi-supervised learning etc.

Experimental Results

- Empirically show that our algorithm converges
- Clustering
- Semi-supervised learning
- MicroRNA data analysis

Experimental Results

Convergence analysis

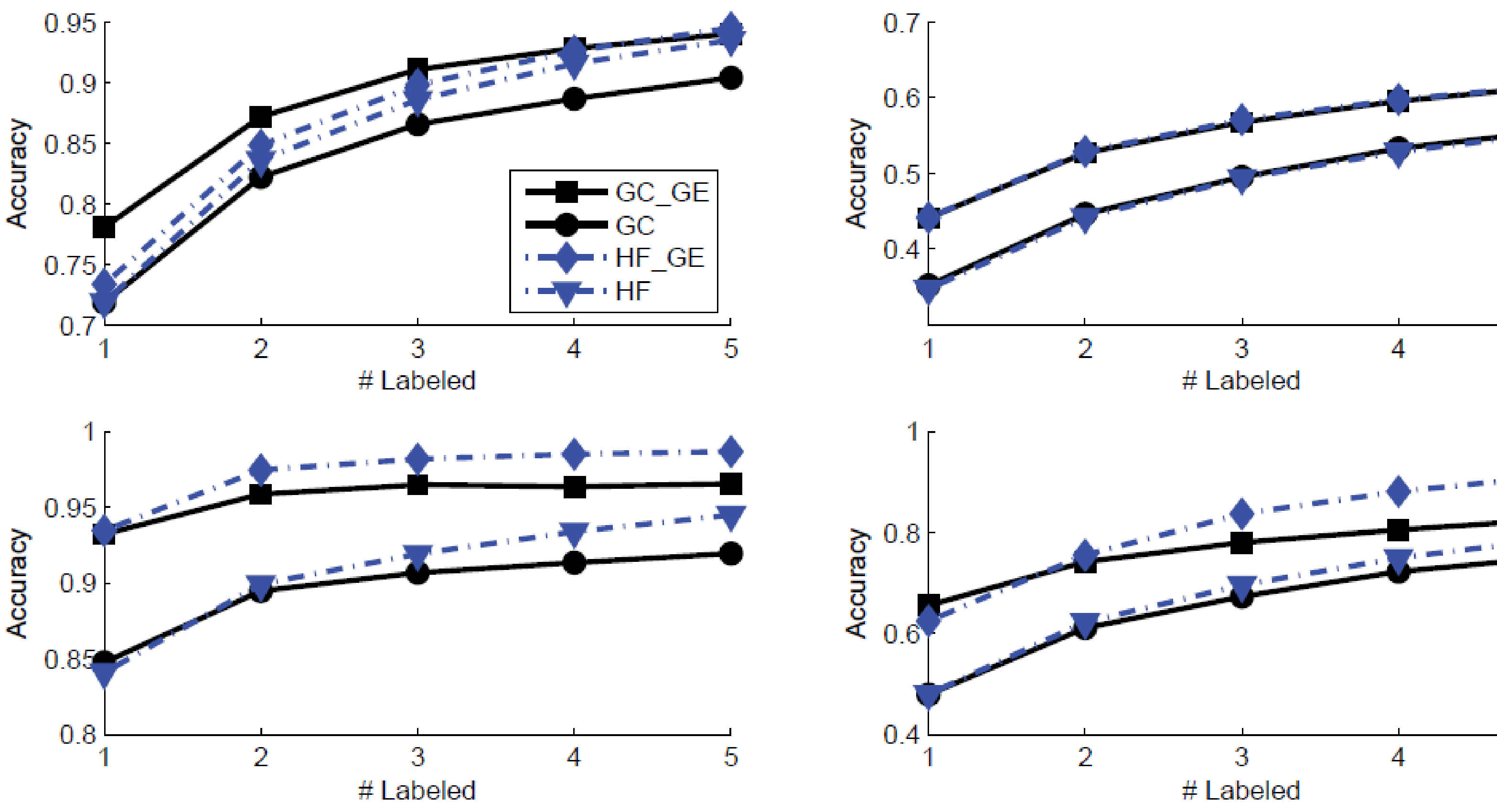


Experimental Results: Clustering

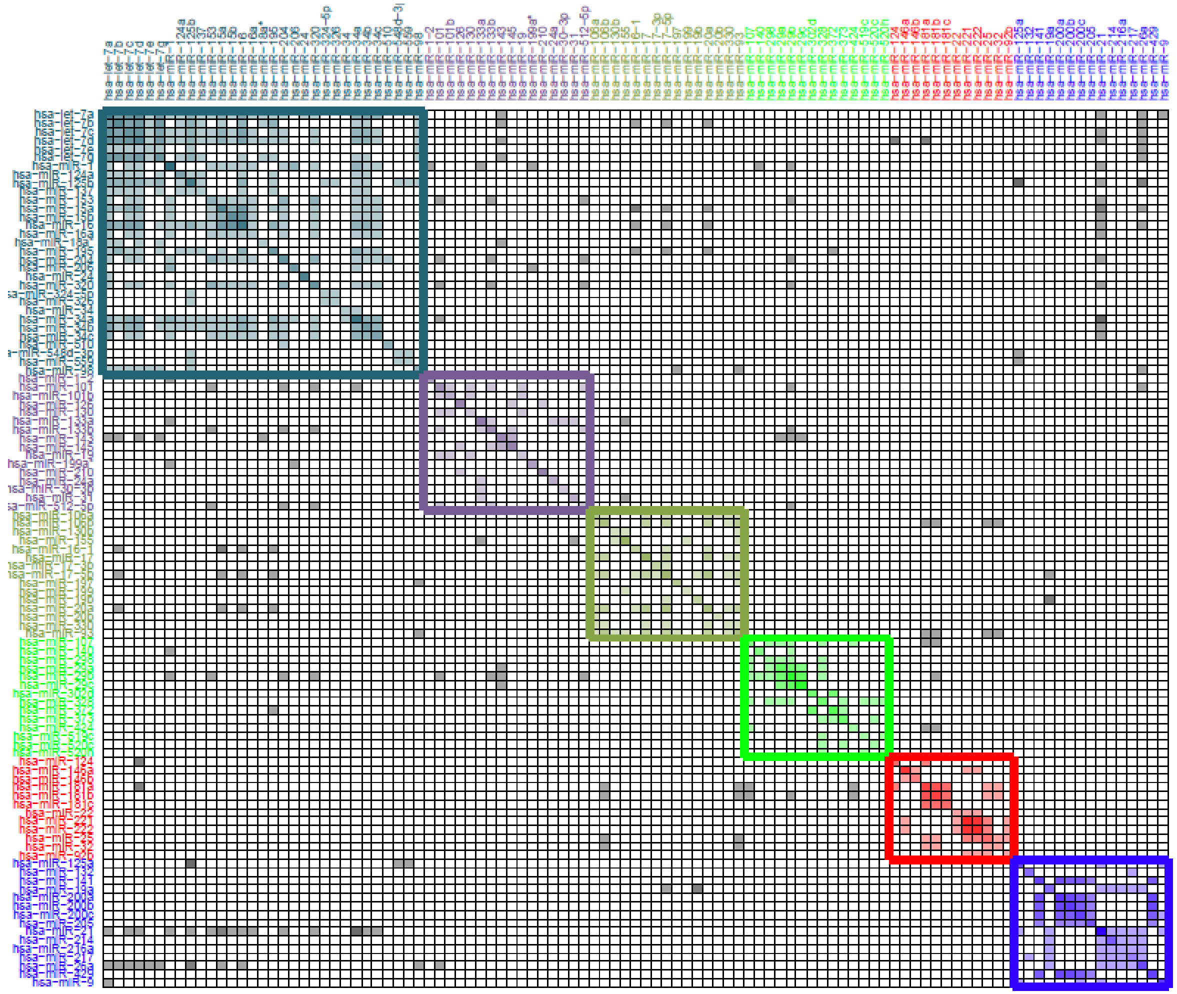
24 UCI Data Sets

	Accuracy				NMI				Purity			
	Km	SC	Ncut	GE	Km	SC	Ncut	GE	Km	SC	Ncut	GE
UMI	0.458	0.471	0.498	**0.644**	0.641	0.646	0.649	**0.763**	0.494	0.505	0.505	**0.667**
COI	0.570	0.614	0.792	**0.839**	0.734	0.750	0.860	**0.879**	0.623	0.658	0.817	**0.840**
ION	0.707	0.702	0.684	**0.880**	0.123	0.193	0.107	**0.446**	0.707	0.730	0.684	**0.880**
JAF	0.744	0.799	0.965	**0.967**	0.809	0.849	0.959	**0.962**	0.774	0.819	0.965	**0.967**
MNI	0.687	0.713	0.820	**0.833**	0.690	0.698	0.748	**0.769**	0.705	0.733	0.820	**0.833**
ORL	0.582	0.683	0.756	**0.775**	0.786	0.834	0.866	**0.891**	0.624	0.713	0.773	**0.802**
PR1	0.716	0.675	0.562	**0.899**	0.129	0.176	0.102	**0.458**	0.726	0.757	0.708	**0.899**
PR2	0.580	0.566	0.569	**0.706**	0.019	0.017	0.013	**0.136**	0.580	0.566	0.569	**0.706**
SOY	0.908	0.871	**1.000**	**1.000**	0.903	0.859	**1.000**	**1.000**	0.924	0.893	**1.000**	**1.000**
SRB	0.480	0.622	**0.699**	0.639	0.232	0.411	**0.454**	0.421	0.512	0.645	**0.699**	0.639
YEA	0.132	0.327	0.302	**0.395**	0.013	0.129	0.126	**0.231**	0.328	0.430	0.436	**0.540**
ZOO	0.264	0.674	0.629	**0.723**	0.116	0.615	0.570	**0.751**	0.423	0.750	0.737	**0.871**
AML	0.688	0.678	0.659	**0.847**	0.100	0.100	0.073	**0.394**	0.696	0.692	0.666	**0.847**
CAR	0.623	0.729	0.719	**0.799**	0.655	0.743	0.738	**0.779**	0.691	0.789	0.788	**0.822**
WIN	0.961	0.936	0.978	**0.983**	0.863	0.845	0.907	**0.926**	0.961	0.943	0.978	**0.983**
LEU	0.879	0.840	0.958	**0.972**	0.559	0.513	0.735	**0.806**	0.879	0.860	0.958	**0.972**
LUN	0.663	0.672	**0.748**	0.704	0.495	0.485	**0.547**	0.473	0.864	0.860	**0.911**	0.828
DER	0.766	0.848	0.955	**0.964**	0.838	0.818	0.905	**0.931**	0.853	0.876	0.955	**0.964**
ECO	0.552	0.496	0.505	**0.631**	0.467	0.458	0.487	**0.549**	0.739	0.770	0.808	**0.851**
GLA	0.452	0.446	0.453	**0.565**	0.320	0.298	0.333	**0.399**	0.549	0.572	**0.652**	0.650
GLI	0.585	0.548	0.559	**0.700**	0.465	0.410	0.398	**0.505**	0.619	0.569	0.601	**0.700**
IRI	0.802	0.746	0.843	**0.953**	0.640	0.514	0.655	**0.849**	0.815	0.758	0.843	**0.953**
MAL	0.911	0.731	0.902	**0.929**	0.569	0.299	0.544	**0.624**	0.911	0.743	0.902	**0.929**
MLL	0.669	0.637	0.687	**0.861**	0.435	0.376	0.426	**0.681**	0.692	0.651	0.687	**0.861**

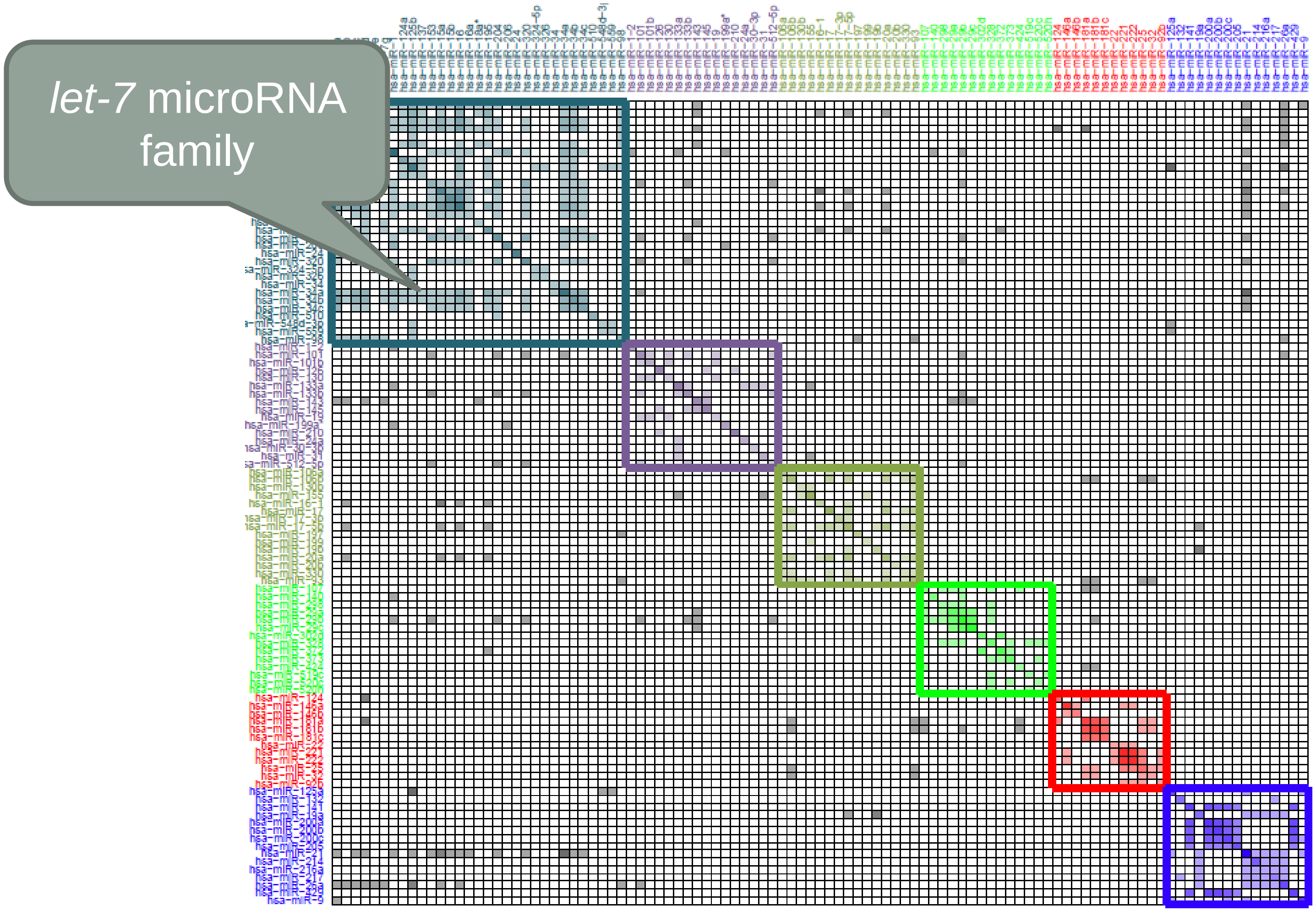
Experimental Results: Semi-supervised Learning



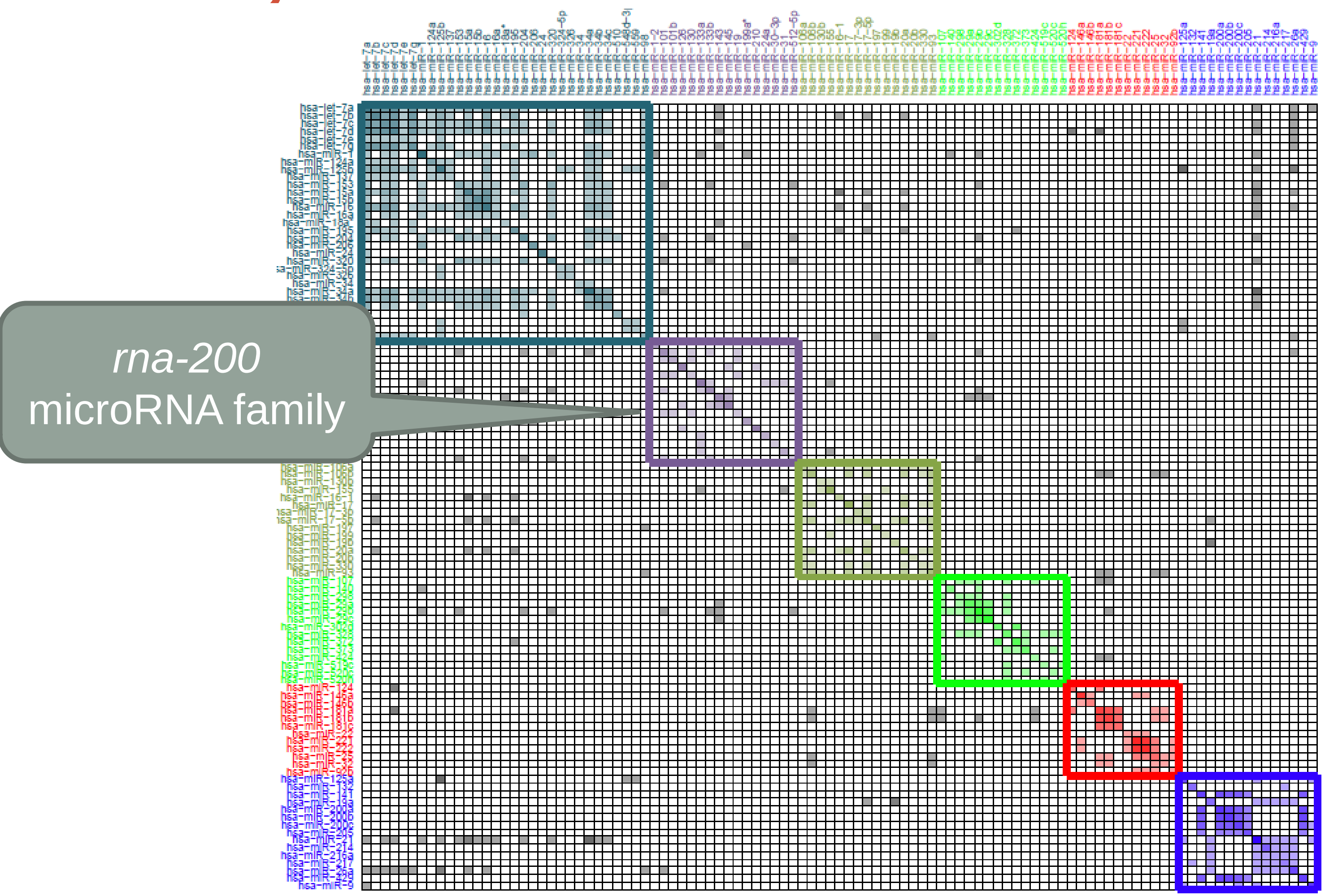
Experimental Results: microRNA function analysis



Experimental Results: microRNA function analysis

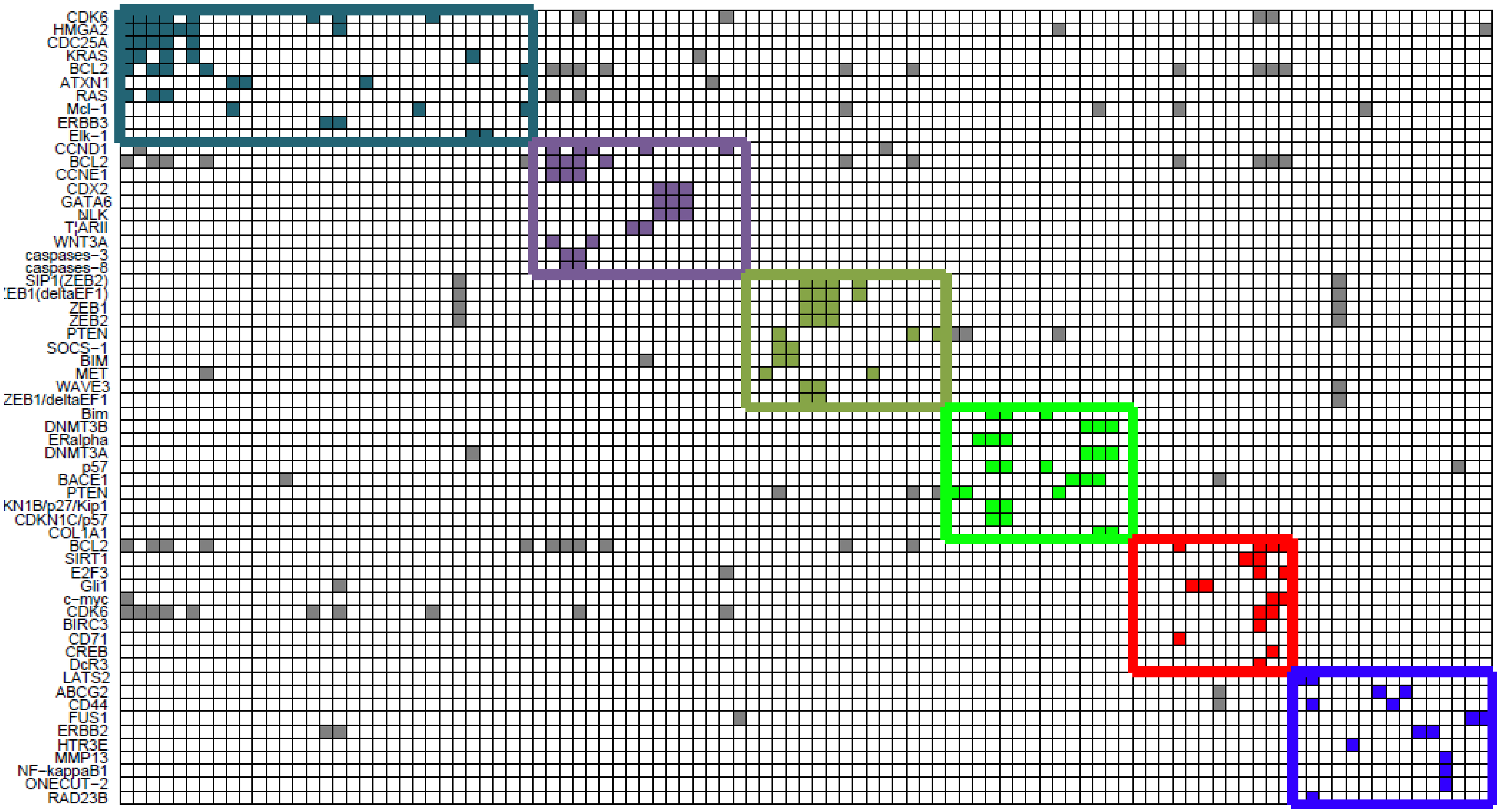


Experimental Results: microRNA function analysis



Experimental Results: microRNA function analysis

- The corresponding genes



Experimental Results: microRNA function analysis

- Observations
  - 6 microRNA groups are identified
  - *let-7* and *mir-200* family a have been reported by other researchers *[Hu 2009, Abbott 2005]*

Conclusions

- A novel social diffusion process model is presented
  - Dynamic graph evolution
  - Analogue of the Mathew effect
- Simple, intuitive, interpretable
  - Directly corresponds to graph language
- Extensive experiments on 24 UCI data sets
  - Better clustering accuracy
  - Better semi-supervised learning performance
- Unsupervised graph-data exploration
  - Almost no parameter
  - Easy to visualize
  - Meaningful results